



# Constituição de *corpus*: critérios, parâmetros e representatividade

## CONCORDANCE

MIGRANTE\_PT

lemma **construir** • 72  
50.15 per million tokens • 0.005%

Get more space

Left context KWIC Right context

1	procura contribuir, na convicção de que a Europa da solidariedade se	<b>constrói</b>	nos momentos mais desafiantes, precisamente como aquele que enfre
2	ações entre mulheres e homens baseada em identidades definidas ou	<b>construídas</b>	social ou culturalmente, enquanto que o sexo é a determinante biológic
3	algo estático ou inato, e adquire um significado social e culturalmente	<b>construído</b>	ao longo do tempo.</s></s> Solicitações baseadas no género podem se
4	rando-se as graves consequências da exclusão, faz-se imprescindível	<b>construir</b>	rigorosas salvaguardas procedimentais no processo de determinação
5	a pelo fato de que as crianças compartilham uma experiência comum,	<b>construída</b>	socialmente, como serem abusadas, abandonadas, empobrecidas ou
6	so de reloca- lização à escala da UE.</s></s> Este sistema deverá ser	<b>construído</b>	com base nos seguintes princípios: uma clara descrição pelos países c
7	usões sociais, afirmar os direitos humanos como motor da cidadania e	<b>construir</b>	uma cidade aberta, diversa e intercultural.</s></s> A não-discriminação
8	ência da opinião pública e da agenda política, podendo contribuir para	<b>construir</b>	, difundir e sustentar preconceitos e estereótipos ligados à migração e
9	lano, previstos na dimensão estratégica e na dimensão operacional; ii)	<b>construir</b>	indicadores de impacto do Plano relativos à integração de migrantes n
10	DS i.</s></s> Analisa os indicadores de resultado do Plano; ii.</s></s>	<b>Constrói</b>	indicadores de impacto por área temática; iii.</s></s> Contacta e/ou reu
11	orização que a cidade e as suas instituições atribuem à diversidade foi	<b>construído</b>	o I Plano Municipal para a Integração de Imigrantes que decorreu de u
12	z-se de um instrumento de planeamento estratégico de base territorial,	<b>construído</b>	através de uma participação conjunta e partilhada e que se estrutura a
13	A1 (Norte/Sul) e aos IC6 e IC7 (Litoral/Interior) e a A13, recentemente	<b>construída</b>	</s></s> A Lei nº 22/ 2012 de 30 de maio veio a estabelecer a reorgani
14	vista a dotar os Grupos de Trabalho de organização e dinâmica foram	<b>construídos</b>	os seguintes instrumentos de trabalho: Análise SWOC e Grelha de Pric
15	erais do Projeto "ODEMIRA INTEGRÁ" Com este projeto, pretende-se	<b>construir</b>	um PMII assente numa participação ativa de toda a comunidade, capa
16	ção do Diagnóstico, com a participação de todos os parceiros.</s></s>	<b>Construíram-se</b>	seis grupos de trabalho por áreas de intervenção e todos os membros

# Constituição de corpus

## Esquema de trabalho

### 1. Constituição do *corpus*

- seleção
- amostragem
- normalização
- codificação
- descrição do *corpus*

# Constituição de corpus

## **Criação do *corpus*: questões iniciais (Sinclair, 1996)**

- Concepção do *corpus*: linguistas vs. especialistas?
  - depende dos objetivos
- Formatos: oportunidades e dificuldades
  - formato digital inicial
  - digitalizações (OCR e interpretação)
  - transcrição (discurso oral e interpretação)

# Constituição de corpus

## Criação do *corpus*: questões iniciais (cont.)

- Permissões e direitos de autor: meios e estratégias
  - depende da utilização: investigação vs. divulgação
    - contratos de licenças (autores e editoras)
    - obras livres de direitos de autor (70-80 anos)  
(não necessariamente edições livres de direitos de autor)
    - autorizações dos entrevistados ou responsáveis

# Constituição de corpus

## Constituição do *corpus* e seleção

- A seleção do que entra para um *corpus* depende dos objetivos de utilização do *corpus*.
  - discurso escrito e discurso oral?
    - *corpus* de referência
    - *corpus* literário
    - *corpus* de especialidade
  - como e onde obter discurso oral?
  - discurso formal e informal? Como e onde obter...

# Constituição de corpus

- A seleção do que entra para um *corpus* depende dos objetivos de utilização do *corpus*.  
(cont.)

- tipologia de textos e amostragem?

- artigos revistas, teses, notícias? Quantos de cada?

- textos típicos/representativos?

- de quê? representatividade estatística vs.  
representatividade de outros elementos

- tipicidade: padrões linguísticos e resultados  
normativos

( → *tipicidade e critérios de amostragem* )

# Constituição de corpus

- A seleção do que entra para um *corpus* depende dos objetivos de utilização do *corpus*. (cont.)
  - período de tempo considerado
  - tamanho (do *corpus*, das subcoleções, das amostras...)
  - amostragem vs. textos integrais (tipos de *corpora*)
  - critérios mínimos, informação e codificação (tipo de texto, autor, sexo, idade, local...)
    - redesenho e 'especialização' do *corpus*

# Constituição de corpus

## Critérios de seleção (gerais)

- **Tipos de texto**

*Corpus* geral/de referência

*Corpus* literário: subgénero (romance, novela, poesia...)

*Corpus* de especialidade: ...?

- artigos
- comunicação técnica interna
- materiais pedagógicos
- materiais divulgação (especialista – especialista;  
especialista – não-especialista)



# Constituição de corpus

## **Critérios de seleção\_(gerais)**

- **Período**

- 20 anos?
- 50 anos?
- ..

- **Língua**

- original?
- tradução?
- língua materna do autor?

# Constituição de corpus

## **Critérios de seleção\_(gerais)**

- **Tamanho**

- das amostras
- do corpus

"The dimension of a corpus are of prime concern for most researchers in the initial conceptualization, and in the public statements. In the long run, they matter very little. The only guidance I would give is that a corpus should be as large as possible and should keep on growing."  
(Sinclair 1991: 18)

Sinclair, J. (1991). "Corpus, Concordance, Collocation", Oxford University Press.

# Constituição de corpus

## **CrITÉrios de seleço (outros)**

- **Publicaço/ediço**
  - local
  - empresa?
  - ...
- **Grau de escolaridade(?)**
  - licenciatura
  - doutoramento
  - ttulo de especialista?

# Constituição de corpus

## Métodos de seleção

- Critérios de seleção normativa (canonicidade/tipicidade/representatividade estatística)
- Critérios de seleção metódica (amostragem da população)
- Critérios de seleção orientados para os resultados (contexto da investigação/constituição do *corpus*)

# Constituição de corpus

## Métodos de seleção (cont.)

- **CrITÉrios de seleção normativa**  
(canonicidade/tipicidade/representatividade estatística)

**Canonicidade:** estatuto de prestÍgio social, cultural, econÓmico que reflete legitimação consensual e normativa.

→ Noção temporal e culturalmente marcada!

# Constituição de corpus

## Métodos de seleção (cont.)

- **CrITÉRIOS de seleção metódica** (amostragem da população)

- amostragem estatística (externa e não interna...)

- externa (composição do *corpus*)

- quantos e quais os textos a selecionar de ente uma subcoleção?

- amostragem aleatória vs.

- amostragem **representativa de diversidade?**

# Constituição de corpus

## Métodos de seleção (cont.)

- **Critérios de seleção orientados para os resultados** (contexto da investigação/constituição do *corpus*)
  - critérios mistos/abordagem pragmática:
    - amostragem
    - tipicidade/canonicidade
    - disponibilidade (formato, direitos de autor, ...)
    - objetivos de investigação (ex.: diversidade...)

# Constituição de corpus

## Constituição de *corpus* e amostragem

- "elegibilidade" (nível 1: critérios de seleção)
- "composição" (nível 2: critérios/intervalos das amostras do corpus/por subcoleção...)

1. de 10% a 50% de textos escritos por doutorados
  2. 10 autores representados por 3 textos
  3. pelo menos 30% de textos relevantes (ex. revistas indexadas) e pelo menos 30% de textos não relevantes
  4. pelo menos 20% de textos curtos (1500 a 6 000 palavras) e pelo menos 20% de textos longos (mais de 80 000 palavras)
- ...

→ *equilíbrio do corpus*



# Constituição de corpus

## Normalização e codificação

‘Data is ontologically different from the world.’

Moisl (2009: 876)

Ex.: um texto digitalizado e em formato eletrónico é diferente do texto original em papel

formatação (tipo de letra, itálicos, maiúsculas, negritos...)

translineação (divisão de palavras por linha)

quebras de página

...

# Constituição de corpus

## Normalização e codificação (cont.)

- A normalização implica decisões a aplicar sobre TODOS os textos sobre
    - formatos
    - alterações
    - codificação de elementos no original
    - codificação de alterações feitas...
- a normalização pode ser mais ou menos relevante

*dados linguísticos?*

# Constituição de corpus

## Normalização e codificação (cont.)

- Exemplo:
  - mantemos/retiramos o número da página?
  - corrigimos erros ortográficos? *voçê?* e *gralhas?* *exmplo?*  
(indicamos essa correção?)
  - mantemos/retiramos os hífens da translineação?
  - mantemos/alteramos a grafia original?

→ *implicações a nível da estatística lexical...*

# Constituição de corpus

## Normalização e codificação (cont.)

- Exemplo:
  - mantemos/identificamos o itálico/negrito/diferentes tipos de letra?
    - *implicações a nível da identificação de neologismos, estrangeirismos, ênfase...*
  - mantemos/normalizamos aspas diferentes?
    - *implicações a nível da identificação de neologismos, estrangeirismos, ênfase, intertextualidade...*

# Constituição de corpus

## Codificação (vs. anotação)

- A **codificação** diz respeito à expressão e marcação de informações inerentes aos textos
  - codificação de elementos ao nível supra textual (cabeçalho do texto; cabeçalho do corpus)
  - codificação de elementos ao nível do texto (parágrafos, formatos, cortes, etc...)
- A **anotação** diz respeito à expressão e marcação de informações extra texto (anotação morfossintática, etc...)

# Constituição de corpus

## Codificação

- CABEÇALHO DO TEXTO *metadados (exemplo CRPC)*

Natureza dos dados

Sigla da natureza dos dados

Fonte

Sigla da fonte

Número de ordem

Nome do autor

Ano de nascimento do autor

Sigla do autor

Nome do jornal/revista

Sigla do nome do jornal/revista

Título

Número do volume

Sigla do título

Nome da disciplina curricular

Sigla da disciplina curricular

Ano de escolaridade

Secção

Número da edição

Número do jornal/revista

Editor

Sigla do editor

Colecção

Sigla da colecção

Localidade da edição

Sigla da localidade

Data

Sigla da data

Género/Tema

Sigla do género/tema

# Constituição de corpus

## Codificação

- CABEÇALHO DO TEXTO *metadados (exemplo CRPC) (cont.)*

Página

Coluna

País da edição

Sigla do país

Directoria

Ficheiro

Número de linhas

Número de palavras

Número de caracteres

Estado

Observações

Introdução

Correcção

Revisão

Data da 1ª edição

País do autor

Língua materna do autor

País de nascimento do autor

Local de nascimento do autor

Endereço web do download

Data do download

Língua do original

Situação de Direitos de Autor

**É possível codificar informação no nome de ficheiro e nome das pastas!**

# Constituição de corpus

## Descrição do *corpus* e codificação

- Elemento essencial para a utilidade e relevância do *corpus*
- Discriminação dos elementos constituintes do corpus, de acordo com os critérios de seleção e amostragem definidos.

"Until we know a lot more about the effects of our design strategies, we must rely on publishing a list of exactly what is in a corpus (...). Users and critics can then **consider the constitution and balance of the corpus as a separate matter from the reporting of the linguistic evidence of the corpus.**"

(Sinclair 1991:13)



# Constituição de corpus

- Critérios de elegibilidade
- Critérios de amostragem
- Normalização
- Codificação e descrição do corpus

**→ permitem a avaliação do corpus!**



Obrigada!



## CONCORDANCE

MIGRANTE\_PT

lemma **construir** • 72  
50.15 per million tokens • 0.005%

Get more space

Left context   KWIC   Right context

	Left context	KWIC	Right context
1	procura contribuir, na convicção de que a Europa da solidariedade se	<b>constrói</b>	nos momentos mais desafiantes, precisamente como aquele que enfre
2	ações entre mulheres e homens baseada em identidades definidas ou	<b>construídas</b>	social ou culturalmente, enquanto que o sexo é a determinante biológic
3	algo estático ou inato, e adquire um significado social e culturalmente	<b>construído</b>	ao longo do tempo.</s></s> Solicitações baseadas no género podem se
4	rando-se as graves consequências da exclusão, faz-se imprescindível	<b>construir</b>	rigorosas salvaguardas procedimentais no processo de determinação
5	a pelo fato de que as crianças compartilham uma experiência comum,	<b>construída</b>	socialmente, como serem abusadas, abandonadas, empobrecidas ou
6	so de reloca- lização à escala da UE.</s></s> Este sistema deverá ser	<b>construído</b>	com base nos seguintes princípios: uma clara descrição pelos países c
7	usões sociais, afirmar os direitos humanos como motor da cidadania e	<b>construir</b>	uma cidade aberta, diversa e intercultural.</s></s> A não-discriminação
8	ência da opinião pública e da agenda política, podendo contribuir para	<b>construir</b>	, difundir e sustentar preconceitos e estereótipos ligados à migração e
9	lano, previstos na dimensão estratégica e na dimensão operacional; ii)	<b>construir</b>	indicadores de impacto do Plano relativos à integração de migrantes n
10	JDS i.</s></s> Analisa os indicadores de resultado do Plano; ii.</s></s>	<b>Constrói</b>	indicadores de impacto por área temática; iii.</s></s> Contacta e/ou reu
11	orização que a cidade e as suas instituições atribuem à diversidade foi	<b>construído</b>	o I Plano Municipal para a Integração de Imigrantes que decorreu de u
12	z-se de um instrumento de planeamento estratégico de base territorial,	<b>construído</b>	através de uma participação conjunta e partilhada e que se estrutura a
13	A1 (Norte/Sul) e aos IC6 e IC7 (Litoral/Interior) e a A13, recentemente	<b>construída</b>	</s></s> A Lei nº 22/ 2012 de 30 de maio veio a estabelecer a reorgani
14	vista a dotar os Grupos de Trabalho de organização e dinâmica foram	<b>construídos</b>	os seguintes instrumentos de trabalho: Análise SWOC e Grelha de Pri
15	erais do Projeto "ODEMIRA INTEGRA" Com este projeto, pretende-se	<b>construir</b>	um PMII assente numa participação ativa de toda a comunidade, capa
16	ção do Diagnóstico, com a participação de todos os parceiros.</s></s>	<b>Construíram-se</b>	seis grupos de trabalho por áreas de intervenção e todos os membros



<b>EN</b>	The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
<b>PT</b>	O apoio da Comissão Europeia à produção desta publicação não constitui um aval do seu conteúdo, que reflete unicamente o ponto de vista dos autores, e a Comissão não pode ser considerada responsável por eventuais utilizações que possam ser feitas com as informações nela contidas.